No specimen left behind: Collections digitisation at the NHM, London*

Vince Smith

Collections for the 21st Century, Florida
5-6 May 2014
Some history...

“the rate of progress by the UK taxonomic institutions in digitising and making collections information available is disappointingly low... there is a significant risk of damage to the international reputation of major institutions such as The Natural History Museum”

House of Lords Science and Technology Committee Report on Taxonomy and Systematics, 2009
Digitisation rates at the NHM (circa 2009)

900 years to digitise the collection!
The prevailing attitude collections digitisation

“However desirable, digitization of all specimens across the globe is a noble but impracticable goal.”

“Digitizing all specimens is not an achievable aim at present”

2010 GBIF Task Group: Global Strategy and Action Plan for the Digitisation of Natural History Collections

Biodiversity Informatics
2010, 7: 120 – 129
More technology, more automation, more speed

Whole drawer scanning

Herbarium sheet scanning

Microscope slide scanning
European collections rising to the challenge

Large-scale data capture & digitisation in France, Netherlands & Finland
A New Voyage of Discovery

Three Focal Areas
1. Scientific discovery
2. Scientific Infrastructure
3. Scientific engagement

Five Challenges
1. The Digital NHM
2. Origins, evolution & futures
3. Biodiversity discovery
4. Natural resources & hazards
5. Science, society & skills

Resources & funding

Measuring success
Digitisation target
20M specimens available by 2017

data.nhm.ac.uk/globe/
NHM collections comprise c.80m objects
Physical register: c.5m
Digital data: 2.8m
Images: 350k

<table>
<thead>
<tr>
<th>Collection area</th>
<th>No of objects</th>
<th>No of type specimens</th>
<th>Physical register</th>
<th>Digital data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Palaeontology</td>
<td>6,919,207</td>
<td>43,146</td>
<td>2,364,232</td>
<td>340,636</td>
</tr>
<tr>
<td>Mineralogy</td>
<td>423,563</td>
<td>615</td>
<td>425,000</td>
<td>402,727</td>
</tr>
<tr>
<td>Botany</td>
<td>5,863,000</td>
<td>172,750</td>
<td>127,200</td>
<td>645,222</td>
</tr>
<tr>
<td>Entomology</td>
<td>33,753,257</td>
<td>612,796</td>
<td>57,197</td>
<td>255,000</td>
</tr>
<tr>
<td>Zoology</td>
<td>27,501,350</td>
<td>325,000</td>
<td>1,986,000</td>
<td>1,160,216</td>
</tr>
<tr>
<td>Library &amp; archives</td>
<td>5,460,000</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>TOTAL</td>
<td>79,920,377</td>
<td>1,154,307</td>
<td>4,959,629</td>
<td>2,803,801</td>
</tr>
</tbody>
</table>
NHM Digital Collections Programme

A 2, 5 and 10 year plan...

To collate, organise and make available one of the world’s most important natural history collections as digital resource, delivering:

1. an online specimen / lot-level database to manage all holdings
2. core meta-data and / or images for key parts of the collection
3. flexible informatics tools

£750,000 for first 2 years
Outline

1. Why
   • Internal objectives & benefits
   • Research opportunity - the iCollections example

2. What
   • How much data to digitise
   • Linking digitisation effort to project benefits

3. How
   • Digi-street pilots, quick wins (herbarium, drawer & slide scanning)
   • Crowdsourcing pilots & options

4. Where
   • NHM Data Portal
   • External Portals (E.g. GBIF, Europeana)

5. Links
   • Crowdfunding
   • H2020 projects (COST, SYNTHESYS, LOD, VRE, Dig. Inf.)
   • Other museums, herbaria & partners (e.g. CETAF & publishers)

6. When
### 1. Why: Objectives

<table>
<thead>
<tr>
<th>Category</th>
<th>2 year</th>
<th>5 year</th>
<th>10 year</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>POLICY &amp; PROTOCOL</strong></td>
<td>Defined data policy and standards</td>
<td>Policies embedded in NHM operating practises</td>
<td>Leaders of process in the digital curatorial world</td>
</tr>
<tr>
<td><strong>DATA CAPTURE</strong></td>
<td>Defined and prioritised collections streams</td>
<td>Portfolio of mass digitisation output projects</td>
<td>Some major collections digitised</td>
</tr>
<tr>
<td><strong>PEOPLE &amp; SKILLS</strong></td>
<td>Mobilised task force, user groups &amp; leadership</td>
<td>Best-practice processes integrated into training</td>
<td>Digital curation as a core part of our practice</td>
</tr>
<tr>
<td><strong>INFRATRUCTURE</strong></td>
<td>Refined collections database, tools &amp; hardware</td>
<td>Future collections database implemented</td>
<td>Broad connections to other large digital collections Crowdsourcing models</td>
</tr>
<tr>
<td><strong>STAKEHOLDERS &amp; GOVERNANCE</strong></td>
<td>Centralised budget, resource &amp; commissioning</td>
<td>Refined metrics and monitoring regime</td>
<td>Proactive engagement of emerging audiences</td>
</tr>
<tr>
<td><strong>PARTNERSHIPS</strong></td>
<td>Crowdfunding, crowdsourcing and partner initiatives piloted</td>
<td>Outsourcing and crowdsourcing and partner initiatives piloted</td>
<td>Leading major international coalitions</td>
</tr>
<tr>
<td><strong>RESEARCH</strong></td>
<td>Research-orientated projects &amp; initiatives</td>
<td>Collaborative research material published</td>
<td>Major contributor to Grand Challenges</td>
</tr>
<tr>
<td><strong>ACCESS</strong></td>
<td>Live Data Portal</td>
<td>Tools, visualisations &amp; analytics</td>
<td>Mass digital collections access for key audiences Integrated global network of users</td>
</tr>
</tbody>
</table>
1. Why: Research opportunity & the iCollections pilot

- Digitisation of British and Irish Lepidoptera collection
- Species poor, specimen rich
- ~500,000 specimens, 5,000 drawers
- Re-curation, imaging, label data, georeferenced
- ~25% complete (started Jan.’13)
- About 50% specimens ‘useable’
- Many specimens in most years (late - 19th century to 1970)
- Provide longer time perspective than most observational records (BMS post-1976)

Using the NHM collections to track long-term seasonal response of butterflies to climate change
1. **Why:** Research opportunity & the iCollections pilot

- 1900-2000, strong correlation between initial collection dates & temperature
- Critical marker on phenological response prior to recent rapid climate change
- Longer time perspective than most observational records (BMS post-1976)
- Museum data available for rare or hard to record species
- An example of unique biological and ecological data from collections

DOI 10.1007/s00484-013-0780-6
2. What: Linking data capture effort to research benefits

### Level of Data Capture

<table>
<thead>
<tr>
<th>Level</th>
<th>Low</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>BM Number of object</td>
<td>Unknown/Not assigned</td>
<td>In label image</td>
<td>Transcribed</td>
<td>Transcribed &amp; linked to digitised register</td>
<td>Transcribed &amp; linked to digitised register</td>
</tr>
<tr>
<td>Object-level identifier if different</td>
<td>Not assigned</td>
<td>In label image, not machine-readable</td>
<td>Transcribed</td>
<td>Machine-readable barcode</td>
<td>Machine-readable barcode</td>
</tr>
<tr>
<td>Location within the NHM</td>
<td>Defined at project level</td>
<td>Collection</td>
<td>Collection + Cabinet</td>
<td>Collection + Cabinet + Drawer</td>
<td>Collection + Cabinet + Drawer + Location within drawer</td>
</tr>
<tr>
<td>Object image</td>
<td>None</td>
<td>Multi-object (e.g. drawer level); not separable</td>
<td>Multi-object (e.g. drawer level); separable</td>
<td>Single object (low res)</td>
<td>Single object (high res)</td>
</tr>
<tr>
<td>Label image</td>
<td>None</td>
<td>Multi-object (e.g. drawer level); human-readable</td>
<td>Single object; human-readable</td>
<td>Single object (e.g. drawer level); transcribed</td>
<td>Single object; transcribed</td>
</tr>
<tr>
<td>Taxonomic name</td>
<td>Unidentified</td>
<td>In label image</td>
<td>Transcribed but above species level</td>
<td>Transcribed to species level</td>
<td>Name linked to current taxonomy</td>
</tr>
<tr>
<td>Type status (if single specimen)</td>
<td>Unknown</td>
<td>In label image</td>
<td>Yes/No</td>
<td>Transcribed &amp; specified</td>
<td>Transcribed &amp; specified</td>
</tr>
<tr>
<td>Geographical location</td>
<td>Unknown in label image</td>
<td>Continent (TDWG level 1)</td>
<td>Country or region (TDWG levels 2-4)</td>
<td>Georeferenced named locality (e.g., town)</td>
<td>Coordinates based on narrative or GPS</td>
</tr>
<tr>
<td>Date of collection</td>
<td>Unknown</td>
<td>In label image</td>
<td>Transcribed</td>
<td>Transcribed month and year or date range</td>
<td>Transcribed date</td>
</tr>
<tr>
<td>Collector</td>
<td>Unknown</td>
<td>In label image</td>
<td>Transcribed</td>
<td>Transcribed &amp; linked to controlled list</td>
<td>Transcribed, linked to controlled list &amp; collector’s notes</td>
</tr>
<tr>
<td>Stratigraphy</td>
<td>Unknown</td>
<td>In label image</td>
<td>Transcribed verbatim</td>
<td>Chronostratigraphic interpretation</td>
<td>Chronostratigraphy, biostratigraphy and lithostratigraphy</td>
</tr>
<tr>
<td>Additional data (e.g. host)</td>
<td>Unknown</td>
<td>Imaged but not transcribed</td>
<td>Partially transcribed</td>
<td>Transcribed in full</td>
<td>Transcribed in full and integrated</td>
</tr>
</tbody>
</table>

### Radar Plots

- **First Sweep**
  - Object Image
  - Collector
  - Date of collection
  - Geographical location
  - Tax. name
  - Stratigraphy

- **Benefit Justified**
  - Coarse biogeography
  - Coarse macroecology
  - Comparative analysis of traits
  - Evolution of disparity

- **Level of Data Capture**
  - Low
  - High

### Research Areas

**Futures**
- Coarse analysis of spp. distribution change
- Coarse Species Distribution Models
- Phenological change

**Resources**
- Semi-automated capture of trait data
- Modelling within-spp variation across range

**Hazards**
- Earliest records of invaders
- Effects of decadal climate oscillations
- Modelling biotic consequences of weather
- Evolution of invasive species
3. How: Digi-street pilots (Herbarium Sheets)
3. **How:** Digi-street pilots (Herbarium Sheets)

33k Specimens per day, 3 shifts (6am-10pm), Netherlands collection complete in 1.5 years

€1.29 Euros per specimen image (if outsourced), transcription at similar cost
3. **How:** Digi-street pilots (Drawer scanning & segmentation)

- SatScan whole drawer scanning
- 30 Million specimens, 130k drawers
- Fast, high res. multi-specimen drawer images (5 mins. each)
- No specimen handling
- Limited drawer / unit tray metadata, plus identifiers
- Specimen segmentation problem
- Digital and physical collection gets out of sync
- Need to automate specimen segmentation
3. **How:** Digi-street pilots (Drawer scanning & segmentation)

Work with Pieter Holtzhausen and Stéfan van der Walt (Stellenbosch University)
Software: Inselect
Main language: Python
3. **How:** Digi-street pilots (Slide scanning)

1. Slides cleaned & barcoded
2. Loaded into hopper (50-100)
3. High resolution scan
4. Images stored & databased

- Originally for histological sections
- Can be adapted for NH specimens
- Many issues:
  - Speed (Max. 500 per day)
  - File size (2-5GB per slide)
  - Network ingestion (100MBps)
  - Reading labels at both ends

- NHM testing 6 systems
- NERC capital grant awarded
- Fully operation early 2015
3. How: Crowdsourcing pilot

- NHM Bird registers
- No advertising
- Hard to transcribe
- Challenging starting project

1 user with 32,629 transcriptions!
92 users with 100+ transcriptions
363 users with 1 transcription
3. How: Crowdsourcing options

Zooniverse Projects

Smithsonian Digital Volunteers

Wikisource transcription (WiR)

Herbarium@Home

Next steps: Survey and review of natural history transcription projects cf. paying transcribers

- A focus for deposition and discovery of NHM research & collections data
- Stable, citable identifiers on datasets & specimen / lot records
- Transparent data quality (un-reviewed, reviewed, reviewed & updated)
- Download (DwCA), web-services & Linked Open Data
- Build using CKAN, with enhanced mapping functionality
4. Where: External Portals

- NHM almost getting data to GBIF!
- Submitting to Europeana portal (via Open-Up)
- Niche collections on Flickr
  - Robust API services
  - Gateway to image analysis projects (e.g. species recognition & trait extraction tools)
5. Links

Crowdfunding
• Personalizes donation
• Scales well
• Requires lots of data
• Most crowdsourcing platforms unsuitable
• Potential for a data visualization to support our needs

H2020 Projects
• EU Research & Innovation funding Programme
• €80 Billion from 2014-2020
• Strong record (EDIT, ViBRANT, SYNTHESYS1/2/3)
• 5 proposals in development for 2014/15
• Better alignment with Digital Collections Programme

Partners
• Major museums & herbaria (Kew, Smithsonian, & Euro.6)
• Umbrella organisations & projects (GBIF, CETAF, iDigBio)
• Universities (e.g. on Image analysis)
• Data publishers (engagement on data & systems)
6. When

Key dates over next 2 years

*Herbarium scanning*
- Pilot – TBC (starting late-2014)

*Drawer scanning*
- Segmentation Software (Aug. 2014)
- Pilots (Ongoing)

*Slide scanner*
- Testing 6 systems (Complete)
- Procurement / purchase (July 2014)
- Pilot projects & system integration (From Sept. 2014)

*Crowdsourcing pilots*
- Draft review paper (Aug. 2014)
- Additional Notes from Nature Project (early 2015)

*NHM Data Portal*
- Internal release (June 2014)
- Public release (Jan. 2015)

*Funding*
Acknowledgements

**Digital Collections Programme**
 Planning: Ian Owens, Ben Atkinson, Dave Thomas, Andy Purvis, Emilie Smith & Vince Smith.

**iCollections**
 Project team: Gordon Paterson, Geoff Martin, Martin Honey, Blanca Huertas, Darrell Siebert, Vladimir Blagoderov, Steve Cafferty, Adrian Hine, Chris Sleep, Mike Sadka, Elisa Cane, Lyndsey Douglas, Joanna Durant, Gerardo Mazzetta, Flavia Toloni, Peter Wing, Malcolm Penn & Liz Duffle.
 Research: Steve Brooks, Angela Self, Flavia Toloni & Tim Sparks.

**Drawer scanning**
 NHM Satscan development: Vladimir Blagoderov, Laurence Livermore & Vince Smith.
 Software: Pieter Holtzhausen & Stéfan van der Walt (Stellenbosch University).

**Slide scanner**
 Testing: Vladimir Blagoderov & Alex Ball.

**Crowdsourcing**
 Pilots (NHM Team): Tim Conyers, Lawrence Brooks & Adrian Hine.
 Review paper: Laurence Livermore & Vince Smith.

**NHM Data Portal**
 Project team: Vince Smith, Darrell Siebert, Dave Thomas & Adrian Hine.
 Development: Ben Scott & Alice Heaton.

*Apologies to anyone I have missed!*